Portfolio Project Option #2

Daniel Luka

MIS581: Capstone – Business Intelligence and Data Analytics

Colorado State University – Global Campus

Dr. Orenthio Goodwin

August 9, 2020

Table of Contents

1. Abstract …………………………………………………………………………….…………......3
2. Introduction…...………………………………………………………………………………..….3
3. Objective.…...……….…………………………………………………………………….........…3
4. Research Hypotheses……..…………………………………..………….………………………..4
5. Overview of Study…...…..……………………...............................................................................5
6. Literature Review…………………………………………………………………………..……...7
7. Research Design…………………………………………………………………………………...8
   1. Methodology……………………………………………………………………………....8
   2. Methods…………………………………………………………………………………...8
   3. Limitations………………………………………………………………………..............9
   4. Privacy Considerations……………………………………………………………………9
   5. Security Considerations………………………………………………………………….10
   6. Ethical Considerations……………………………………………………………...........11
8. Findings…………………………………………………………………………………………..11
9. Conclusion…………………………………………………………………………………….….14
10. Recommendations………………………………………………………………………………..15
11. References……………………………………………………………………………………..…16

Global Tobacco Prevalence Reduction Initiative Efficacy

**1 Abstract**

While on the decline, the seven million deaths per year that are associated with tobacco demand attention and mitigation (WHO, 2017). This study attempts to leverage tobacco prevalence and reduction initiative (MPOWER) data from the World Health Organization (WHO), to determine the relationship between initiative implementation and prevalence. This quantitative analysis employs descriptive and predictive statistics to that end. Results indicate weak and varying positive correlations and r-squared values between MPOWER initiatives and prevalence, which runs counter to some country and region-specific perspectives. Further, while overall MPOWER participation is increasing over time, time itself has a moderately strong positive association with participation.

**2 Introduction**

Globally, tobacco is associated with circa seven million deaths per year (WHO, 2017). Further, tobacco consistently maintains its position as the prominent cause of premature death within the US (Owens et al., 2020). In fact, the US Preventative Services Task Force (USPSTF) recommends mitigative intervention measures to reduce the incidence (Owens et al., 2020). Accordingly, the following analysis will focus on the current state of utilization and mitigation techniques.

**3 Objective**

The dataset in question, from the world health organization (WHO), is country-specific tobacco initiative and utilization incidence (WHO, n.d. c). Therefore, the objective of the research is to determine the relationships between the monitor, protect, offer, warn, enforce, and raise taxes components of the MPOWER framework and tobacco use prevalence (WHO, 2017). This aim will result in a distinction of the high and low performing components, which will help low participating countries prioritize to improve their prevalence. Worst-case scenario, none of the MPOWER initiatives are helpful and countries can prioritize other methods of population tobacco cessation. Alternatively, the WHO could use this feedback to derive new initiatives or increase the efficacy of those existing.

**4 Research Hypotheses**

The primary business question is: Do the WHO MPOWER indicators influence tobacco prevalence? For this first hypothesis, the null hypothesis would posit that none of the indicators influence tobacco prevalence, where the alternate hypothesis would be accepted if any indicator had a significant influence on prevalence. This question is relational in nature, as it is concerned with the impact of attributes on one another (CSU Global, n.d. b). This alternate hypothesis is complex and associative in nature, given the inclusion of numerous variables and relational focus (CSU Global, n.d. a).

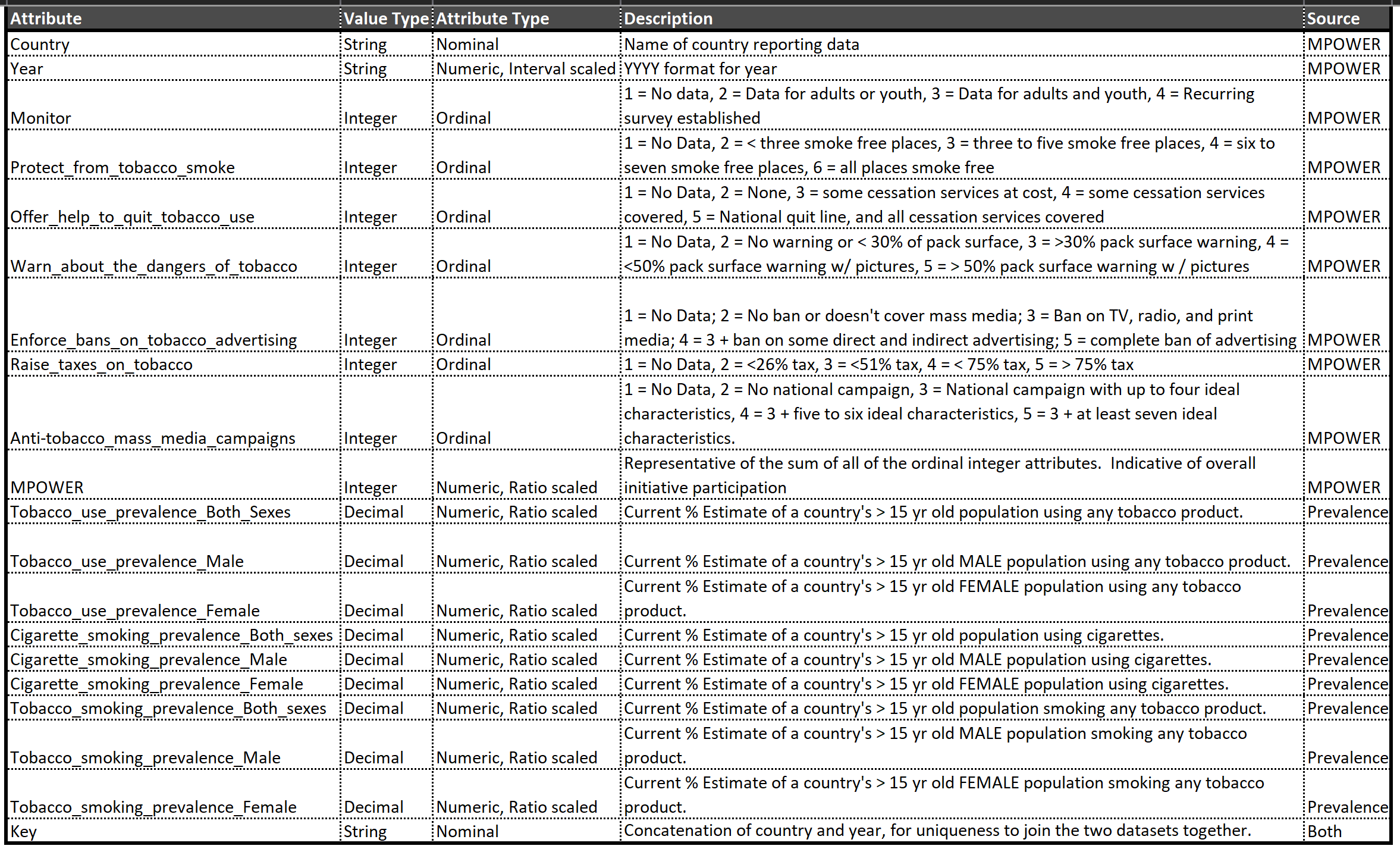
Following that important high-level relational question, a comparative follow-up question is in order (CSU Global, n.d. b). How comparable is the efficacy of each specific MPOWER initiative as they relate to tobacco prevalence? The second hypothesis is complex as it has multiple MPOWER initiatives as independent variables (CSU Global, n.d. a). The null hypothesis here would imply that no initiative maintains a comparative advantage over another.

Given the yearly grain of this country specific data, a trending research question is in order (CSU Global, n.d. b). Has MPOWER initiative participation improved over time? The complex, directional hypothesis would posit that participation has increased over time (CSU Global, n.d. a). This alternate hypothesis is attractive, as the WHO is encouraging and tracking progress towards the initiatives. A null hypothesis acceptance would result from flat initiative participation.

**5 Overview of Study**

With the primary goal of determining the efficacy of MPOWER initiatives on tobacco prevalence, this study will shed light program impact, utilizing the WHO’s own data. Founded circa 1948, post the formation of the United Nations in 1945, the WHO’s current 194 member states and 7,000 employees aim to increase health of individuals across the globe (WHO, n.d. b). Although selecting a US-based organization would provide insights that are particularly relevant for me as a US citizen, the potential comparative insights allowable with multi-country data are enticing. The WHO is a global healthcare leader that develops and tracks progress towards priorities that impact the health of its membership populations (WHO, n.d. d). The financial picture of the WHO is unique in that it is reliant on donations, both within and outside of the context of its member states, to provide a substantial investment of approximately three billion per year to fund its strategy and operations (WHO, n.d. d). Deriving and monitoring strategic global health objectives necessitates consolidated data. The WHO’s data repository solution, referred to as the Global Health Observatory (GHO), hosts data for over a thousand priority health indicators (WHO, n.d. a).

Specifically, it employs an amalgamation of country and year level tobacco utilization and prevention data. The prevention component is sourced from the MPOWER data, which provides a sliding numeric scale ascending from least to highest commitment, to describe the countries level of tobacco monitoring, utilization protection, assistance offering, danger warning, advertising ban, tax increase, and anti-tobacco mass media campaigns (WHO, n.d. c). The utilization component is sourced from prevalence data set, which provides gender-specific percentage incidence estimations for cigarette smoking, tobacco smoking, and tobacco utilization, aggregated by country and year (WHO, n.d. c). In total, the dataset is representative of n = 1,366 observations that have 19 attributes, though only n = 894 observations from the MPOWER set had corresponding prevalence information. An overview of the available attributes is displayed in figure one below, the project data dictionary.



*Figure 1* Data Dictionary. Adapted from WHO. (n.d. c). GHO | By category | Tobacco control. Retrieved from <https://apps.who.int/gho/data/node.main.Tobacco?lang=en>

Generally, the capstone portfolio project dictated analysis of a public dataset belonging to an organization of choice (CSU Global, n.d. b). Given the open-ended nature of the project, the selection was determined based on my personal, professional, and academic experiences. Throughout my analytic tenure, my focus has remained exclusively on the healthcare industry. Accordingly, the scope of the requisite analytic task will be limited to that of healthcare. Primarily, tobacco utilization is a personal interest of mine, given previous personal and family incidence. The importance is further justified from tobacco’s position as the prominent cause of premature death within the US (Owens et al., 2020). With the aforementioned data providing the foundation to test the established hypotheses, a literature review is in order to confirm current direction, prior to discussing research design and results (CSU Global, n.d. c).

**6 Literature Review**

Importantly, the evidence-based MPOWER framework intends to reverse the current and unfortunate incidence of tobacco utilization worldwide (Ahluwalia, Arrazola, & Graffunder, 2017). Regrettably, there is minimal existing evidence to supporting the efficacy of said framework (Ngo, Cheng, Chaloupka, & Shang, 2017). To that end, Ahluwalia et al. (2017) highlight the incidence reduction in Uruguay as a result of it implementing a leading number of MPOWER initiatives. While this country-specific example is interesting, it is anecdotal.

A more comprehensive analysis is presented by Heydari, Zaatari, Al-Lawati, El-Awa, and Fouad (2018), where they focus on 22 countries that comprise the Eastern Mediterranean Region. Heydari et al. (2018) were generally more concerned with adoption of MPOWER initiatives and found that 68% of their countries experienced an increase in MPOWER adoption. The focus on adoption in lieu of efficacy is a gap. Another region, Europe, is analyzed by Ngo et al. (2017), where they employ Euromonitor smoking prevalence and consumption data, in addition to world bank demographic data. Notably, they totaled the score of all MPOWER initiatives for a consolidated metric and demonstrated a corresponding increase with prevalence and consumption reduction (Ngo et al., 2017).

Finally, a consolidated review of the MPOWER efficacy across all countries through 2014 was done by Levy, Yuan, Luo, and Mays (2018). Their study employed the same WHO prevalence data as this research project, though they used it to calculate a relationship between the initiatives and smoking attributable deaths (Levy et al., 2018). Although they demonstrated a reduction in deaths as a result of adopting the highest level of each MPOWER initiative, they don’t analyze the intermediary layers (Levy et al., 2018). Therefore, the current research project will help to fill the perceived research gap, by specifically focusing on the impact of MPOWER initiatives and tobacco prevalence across all countries.

**7 Research Design**

**7.1 Methodology**

This research project is driven by an empirical epistemology, meaning it is in search of observable insights (O’Leary, 2017). Essentially, it is relying on the analysis of the available data attributes to prove or disprove each hypothesis. More specifically, this approach is positivist, given its employment of the scientific method through hypothesis testing (O’Leary, 2017). Lastly, the research is performed with a quantitative approach, as it is almost entirely reliant on numerical attributes (O’Leary, 2017).

**7.2 Methods**

Concerning specific analytic methods used, they cover the areas of description, prediction, and prescription (Sharda, Delen, & Turban, 2018). To begin with SAS, descriptive inclusions are summary statistics and correlation through proc corr (Elliot & Woodward, 2016). A correlation matrix of MPOWER attribute participation and tobacco prevalence will be employed as a relational visualization. From there, regression will verify the relational potential established with the correlation matrix. Regression will be further supported by proc r-squared, which will attempt to show the most advantageous combination of attributes for r-squared values. For visualizations, geographical data will be displayed with the map functionality of Tableau. Also, scatter plots will visualize the relationship between MPOWER initiatives and prevalence. Finally, trended MPOWER participation and prevalence can be visualized.

**7.3 Limitations**

Resoundingly, the largest limitation revolves around the relatively low number of observations (n=1,366). Unfortunately, the data has only been collected six times from 2007 to 2018, so each country has six rows at most. This concern is slightly amplified when further limiting the observations to those that have corresponding tobacco utilization info (n= 894). This problem will continue to be mitigated with time as new surveys are completed.

Other potential limitations and challenges were considered relating to the non-U.S. centricity of the data in question. Largely, it was determined this dataset is absent of any international specific limitations or challenges. First, the terms have been standardized across countries by the WHO, who is both the deriver of definitions and collector of data. Second, the available programming languages and tools are utilized around the world and on domestic and international data alike. Third, the data is mostly quantitative in nature, lending itself well to this specific research study. Essentially, percentages and standardized nominal and ordinal categories make sense from the perspective of any of the countries contained therein.

**7.4 Privacy Considerations**

From the privacy perspective, the protection and de-identification of this type of information is guaranteed by the HIPAA legislation of 1996 (White & Ariyachandra, 2016). Thankfully, the grain of the sourced WHO tobacco MPOWER initiative and utilization dataset ensures privacy. Observations within this dataset are at the country and year level, which means the records cannot be linked to the individual level. This is important, as linkage potential is a privacy concern surrounding the possibility of increasing the identifiability of a dataset by joining it to another (Salerno, Knoppers, Lee, Hlaing, & Goodman, 2017). As there are no identifiers below the level of country and year, the linkage potential for the WHO dataset is adequately limited. The confidentiality of underlying patient information is protected through the level of aggregation that occurred to present the data in the current state (White & Ariyachandra, 2016). Another important privacy consideration is that of consent, which is a process that provides a sense of ownership to the subjects who are represented in the data by allowing them to opt-in or out of data collection or aggregation (White & Ariyachandra, 2016). Consent in this study is implied, due to the voluntary nature of submission of member country data.

**7.5 Security Considerations**

Data security during data analytics is primarily concerned with safeguarding the data from outside access (White & Ariyachandra, 2016). Thankfully, there are numerous computer security software applications that focus on intrusion detection and prevention, malware detection, risk management, firewalls, log management, and security information and event management (SIEM) (Rassam, Maarof, Zainal, 2017). Logically, the level of security has a direct relationship with the identifiability of the information it is to protect. Also relevant to the security consideration is where the data will be stored. Throughout this capstone project, data will be stored and analyzed on my local machine, GitHub, SAS VM, and Tableau.

**7.6 Ethical Considerations**

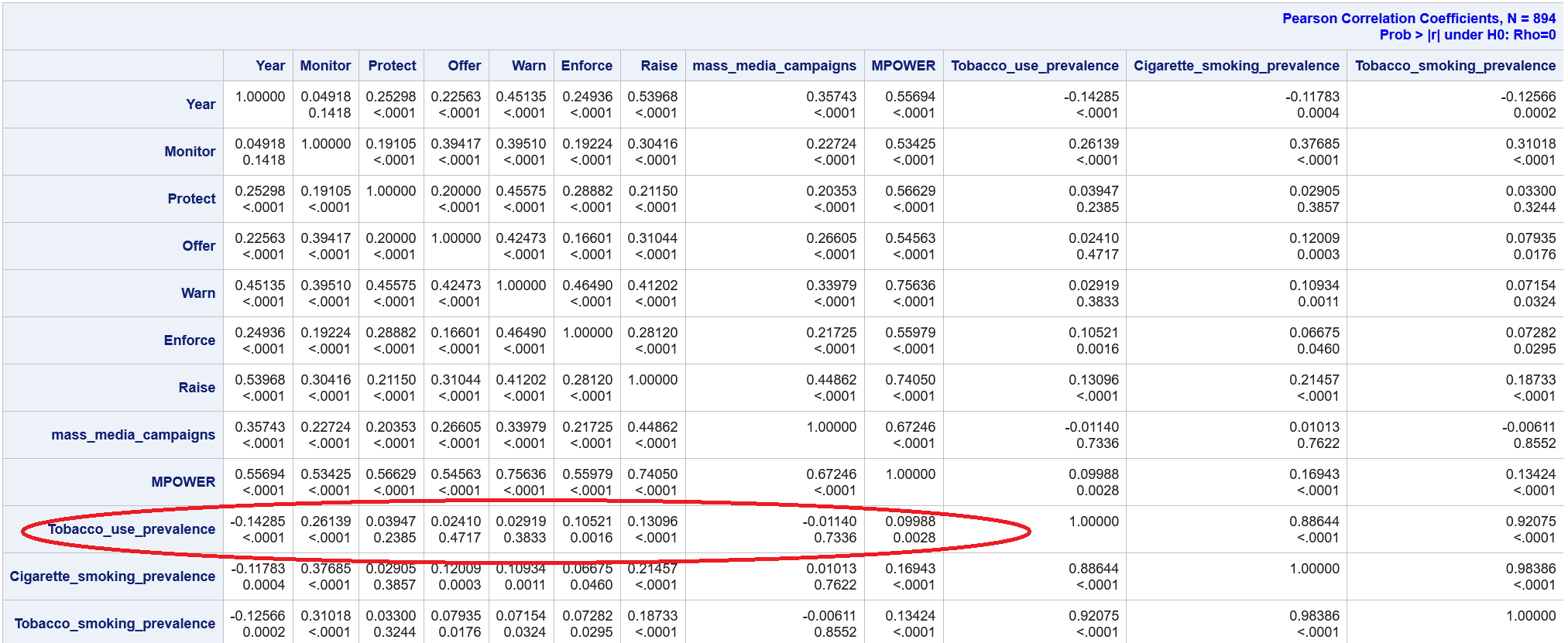
Ethical considerations exist within and outside of the scope of an overall security and privacy framework. While sourcing data, the utilization and source need to be studied for appropriateness (White and Ariyachandra, 2016). The utilization of this information is strictly educational, which is a justifiable use case. Further, the WHO is a well-known, reputable, and ultimately reliable source of information. The credibility of the WHO is favorable to the academic pursuit and integrity of the supporting and resulting data. Also, the publicly available nature of the dataset avoids potential access ethical considerations (Open Data Institute, n.d.). Another important ethical consideration deals with the effect the analysis will have on people, both positive and negative (Open Data Institute, n.d.). The hopefully positive impact is guided by the research project aim of determining the relationship between MPOWER initiatives and tobacco prevalence. The desired outcome of tobacco prevalence reduction would be a net positive to society, with limited to no potential negative impact. Ultimately, the potential positive or negative impact is further limited by the small intended audience of this analysis. The extent of data sharing is an important ethical consideration in and of itself, though not for the narrow scope this academic research project (Open Data Institute, n.d.).

**8 Findings**

Straight away, the correlation matrix resulted in perplexing findings. The MPOWER initiatives showed little to moderate positive correlation with tobacco prevalence. Monitor had the most significant correlation, at .26 (p<.001), followed by enforce and raise at .1 and .09, respectively. These results are curious as they suggest an increase in MPOWER initiative performance ultimately has no significant bearing on prevalence, which suggests the acceptance of the null hypothesis for the first hypothesis. Results were even less advantageous when comparing prevalence to the aggregate MPOWER score, with weak correlation (.09988 p<.001) and predictive value (r-squared of .01). Additionally, the positive nature of the correlations signifies that an increase in the MPOWER initiative ends up increasing prevalence, although only marginally so.

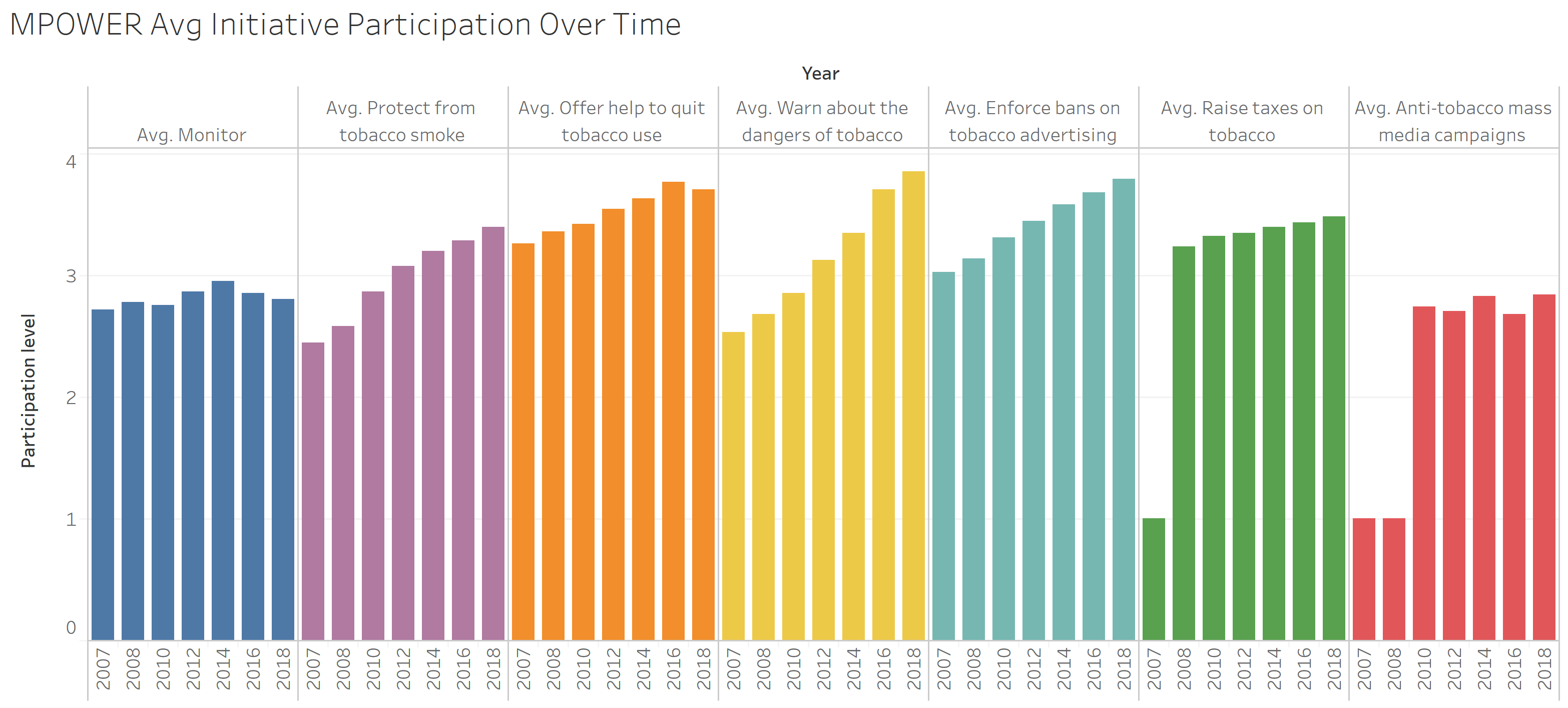
Continuing the analysis of the first business question, regression was employed for the dependent variable of tobacco prevalence and the independent variables of the mpower initiatives and year. Prevalence vs monitor was the first regression, given it had the strongest identified correlation. Conceptually, it makes sense that as monitoring increases so too would identified prevalence. The resulting r-squared of .0683 serves as confirmatory to the correlation analysis, as only 7% of the movement in prevalence is explainable by the monitor attribute. The remaining r-squared for protect, offer, warn, enforce, and raise were .0016, .0006, .0009, .0111, and .00172, respectively. Surprisingly, the r-squared for enforce was greater than that of monitor, indicating it has a greater predictive potential, though minimally so at best. Next, proc r-squared was utilized to determine potentially effective combinations of the MPOWER attributes vs prevalence. Here, a drop-off occurred after the concatenation of three attributes, namely year, monitor, and raise with an r-squared of .1208. Including all MPOWER initiatives and year resulted in an r-squared of .1554, meaning ~78% of the variation in the prevalence cannot be associated with the currently included attributes. Importantly, the resulting regression fit plots and scatter plots on Tableau confirm the variability and unpredictability of prevalence across MPOWER initiative participation values.

Within the second business question, results indicate an acceptance of the alternative hypothesis, as the MPOWER initiatives have varying degrees of correlation and predictive capacity. The protect, offer, and warn components are clustered closely together, though the remainder have substantial variation. Regression results again served as confirmatory here, though quite a few were similar in their level of insignificance. Figure two below is the resulting correlation matrix.



*Figure 2* Correlation Matrix. Adapted from WHO. (n.d. c). GHO | By category | Tobacco control. Retrieved from <https://apps.who.int/gho/data/node.main.Tobacco?lang=en>

Similar to the second, the third business question results imply an acceptance of the null hypothesis. Here, the correlation matrix indicates year has a moderately strong positive correlation with MPOWER at 0.55694 (p<.001). Additionally, the r-squared value of year as it relates to MPOWER participation is .3102, meaning third of the variance in MPOWER score is explainable by year. Essentially, the alternative hypothesis can be accepted as MPOWER participation has a significant, positive relationship with time. Unfortunately, the r-squared results indicate that year alone is not a sufficient predictor of MPOWER participation. Tangentially related, prevalence has decreased over time, though not significantly so. So, while prevalence and participation are moving in the right direction, a causational relationship cannot be determined per the acceptance of the null hypothesis for the first business question. Figure three below displays the progress in average participation over time.



*Figure 3* MPOWER Performance Over Time. Adapted from WHO. (n.d. c). GHO | By category | Tobacco control. Retrieved from <https://apps.who.int/gho/data/node.main.Tobacco?lang=en>

**9 Conclusion**

So far, initial descriptive results are indicating acceptance of the alternative hypothesis for the second business question, as the MPOWER initiatives display varying levels of impact, though none are particularly strong. The alternative hypothesis is also accepted for the third business question, given the moderately strong and significant correlation. The Null hypothesis was only accepted for the first business question. For the primary business question, this means that no MPOWER initiative demonstrated significant relationship to tobacco prevalence. Importantly, while prevalence decreases by year overall, the corresponding negative correlation and r-squared is weak. Therefore, the implications of this analysis are gloomy for the WHO MPOWER initiatives. On the bright side, MPOWER participation has increased significantly over time. Still, the analysis was not able to demonstrate a significant impact of MPOWER initiative on performance, which runs counter to the established understanding, so further validation through reproduction and incremental analysis is suggested. If validated, the MPOWER initiatives necessitate reconsideration to improve their efficacy.

**10 Recommendations**

Within the context of future research, there are a few incremental considerations. Given the exceedingly low r-squared potential with the included attributes, additional information should be sourced for future iterations. Notably absent from the dataset in question are country-specific demographics and social determinants of health, which are likely contributing to the prevalence. Additionally, sourcing data at a level of granularity lower than country (state, county) would allow for more actionable local insights and increased number observations.

**References**

Ahluwalia, I. B., Arrazola, R. A., & Graffunder, C. (2017). Measuring progress in tobacco prevention and control: the role of surveillance.

CSU Global. (n.d. a). Module 01: The Components of Research Project Design. In MIS583: Capstone Business Intelligence and Data Analytics: Spring D 2020 [Class lecture slide]. Retrieved from <https://csuglobal.instructure.com/courses/24754/modules/items/1193243>

CSU Global. (n.d. b). Module 02: Business Questions. In MIS542: Business Analytics: Spring D 2019 [Class lecture slide]. Retrieved from <https://csuglobal.instructure.com/courses/10900/modules/items/530463>

CSU Global. (n.d. c). Module 05: Critical Thinking. In MIS581: Capstone – Business Intelligence and Data Analytics: Spring D 2020. Retrieved from <https://csuglobal.instructure.com/courses/24754/assignments/499134>

CSU Global. (n.d. d). Module 08: Portfolio Project. In MIS581: Capstone – Business Intelligence and Data Analytics: Spring D 2020. Retrieved from <https://csuglobal.instructure.com/courses/24754/assignments/499150>

Elliott, A. C., & Woodward, W. A. (2016). SAS essentials: Mastering SAS for data analytics (Second ed.). Hoboken, NJ: Wiley.

Heydari, G., Zaatari, G., Al-Lawati, J. A., El-Awa, F., & Fouad, H. (2018). MPOWER, needs and challenges: trends in the implementation of the WHO FCTC in the Eastern Mediterranean Region. Eastern Mediterranean Health Journal, 24(1).

Levy, D. T., Yuan, Z., Luo, Y., & Mays, D. (2018). Seven years of progress in tobacco control: an evaluation of the effect of nations meeting the highest level MPOWER measures between 2007 and 2014. Tobacco control, 27(1), 50-57.

Luka, D. (2020 a). Github Portfolio Repository. Retrieved from <https://github.com/danielluka/MIS581-Portfolio-Project-Option-2>

Luka, D. (2020 b). Tableau Profile. Retrieved from <https://public.tableau.com/profile/daniel.luka#!/>

O’Leary, Z. (2017). The Essential Guide to Doing Your Research Project (3rd ed.). Thousand Oaks, CA: SAGE Publications Inc.

Open Data Institute (ODI). (n.d.). The data ethics canvas – The ODI. The ODI – Open Data Institute. Retrieved July 16, 2020, from <https://theodi.org/article/data-ethics-canvas/>

Owens, D. K., Davidson, K. W., Krist, A. H., Barry, M. J., Cabana, M., Caughey, A. B., ... & Kubik, M. (2020). Primary care interventions for prevention and cessation of tobacco use in children and adolescents: US preventive services task force recommendation statement. Jama, 323(16), 1590-1598.

Ngo, A., Cheng, K. W., Chaloupka, F. J., & Shang, C. (2017). The effect of MPOWER scores on cigarette smoking prevalence and consumption. Preventive medicine, 105, S10-S14.

Rassam, M. A., Maarof, M. A., & Zainal, A. (2017). Big Data Analytics Adoption for Cybersecurity: A Review of Current Solutions, Requirements, Challenges and Trends. Journal of Information Assurance & Security, 12(4), 124–145.

Salerno, J., Knoppers, B. M., Lee, L. M., Hlaing, W. M., & Goodman, K. W. (2017). Ethics, big data and computing in epidemiology and public health. Annals of Epidemiology, 27(5), 297-301.

Sharda, R., Delen, D., & Turban, E. (2018). Business intelligence, analytics, and data science: A managerial perspective (4th ed.).

White, G., & Ariyachandra, T. (2016). Big Data and ethics: examining the grey areas of big data analytics. Issues in Information Systems, 17(4).

WHO. (n.d. a). About the Observatory. Retrieved from <https://www.who.int/data/gho/info/about-the-observatory>

WHO. (n.d. b). About WHO. Retrieved from <https://www.who.int/about>

WHO. (n.d. c). GHO | By category | Tobacco control. Retrieved from <https://apps.who.int/gho/data/node.main.Tobacco?lang=en>

WHO (n.d. d) Invest in WHO to invest in Health for All. Retrieved from <https://www.who.int/about/planning-finance-and-accountability/financing-campaign>

WHO (2017). WHO report on the global tobacco epidemic, 2017: monitoring tobacco use and prevention policies. World Health Organization.